INF 558: Building Knowledge Graph

Homework 4: CRF

Q.1) What was the information you were looking for? Describe the labels you chose and why. Also include at least one screenshot of the webpage you are using and show where is the information you are looking for.

My source of unstructured data is from indeed.com, which is a job portal. Also my previous assignments retrieved structural data and our project is based on scraping from job portals, hence this selection.

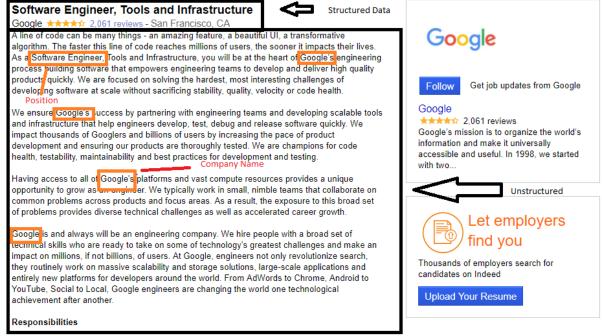
I have used scrapy to crawl the job posting and retrieve the unstructured job description.

Scrapy spider: indeed.py

I have stored the output of this crawled data in a json file: Abhishek Dhameja hw4.json.

The following are the key attributes I planned to retrieve from the unstructured data:

- 1. Company Name (C)
- 2. Position (P)
- 3. Skills (S)



• Lead/contribute to engineering efforts from design to implementation, solving complex

Responsibilities

- Lead/contribute to engineering efforts from design to implementation, solving complex technical challenges around developer and engineering productivity and velocity
- . Design and build advanced automated build, test and release infrastructure
- . Drive adoption of best practices in code health, testing, and maintainability
- Analyze and decompose complex software systems and collaborate with cross-functional teams to influence design for testability

Skills

Qualifications Minimum qualifications

- BS in Computer Science or related technical field or equivalent practical experience
- Software development experience in one or more general purpose programming languages
- Experience in at least one of the following: test automation, refactoring code, test-driven development, build infrastructure, optimizing software, debugging, building tools and testing frameworks

Preferred qualifications:

- Master's or PhD in Computer Science or related technical field
- Experience with one or more general purpose programming languages including but not limited to: Java, C/C++, C#, Objective C, Python, JavaScript, or Go
- Scripting skills in Python, Perl, Shell or another common language

Q.2) What kind of tags did you tag your data with? Explain your choices.

Following tags are collected from the Job description which are scraped from the website:

- 1. Company Name (tagged as "C")
- 2. Position (tagged as "P")
- 3. Skill (tagged as "S")
- 4. Irrelevant data (tagged as "I") (the data which do not belong to any of the above three categories)

I have manually marked 50 training records and 20 test records with C/P/S tag. Build the CRF model on the train data.

Predicted and compared with the manually tagged test data.

Q.3) Report your classifier's precision, recall and F-1 measure. Why did your classifier perform well (or not satisfactorily)?

Find the snapshot of the classifier's performance:

Process finished with exit code 0

- 1. Classifier performs decently in case of the P and the S tag, as all the jobs did mention in the description the position and the skill.
- 2. It performs poorly in case of the C tag as most of the company have not mentioned their names in the description(as already present in the structured heading)(We can retrieve that information from the heading)
- 3. The classifier's performance is boosted as I have used POS tags as one of the features among others